The Blind Estimation of Reverberation Time

Rama Ratnam,[1,6] Douglas L. Jones,[1,2,3,6] William D. O'Brien Jr.,[1,2,6]
Charissa Lansing,[1,4,6] Robert C. Bilger,[4,6] and Albert S. Feng[1,5,6]

[1]*Beckman Institute for Advanced Science and Technology*

[2]*Department of Electrical & Computer Engineering*

[3]*Coordinated Science Laboratory*

[4]*Department of Speech and Hearing Science*

[5]*Department of Molecular & Integrative Physiology*

[6]*University of Illinois at Urbana-Champaign*

(Dated: 6th February 2003)

The reverberation time (RT) is an important parameter for characterizing the quality of an auditory space. Sounds in reverberant environments are subject to coloration. This affects speech intelligibility and sound localization. Many state-of-the-art audio signal processing algorithms, for example in hearing aids and telephony, are expected to have the ability to evaluate the characteristics of the listening environment and turn on an appropriate processing strategy accordingly. Thus, a method for characterization of room RT based on passively received microphone signals represents an important enabling technology. Current RT estimators, such as Schroeder's method or regression, depend on a controlled sound source, and thus cannot produce an online, blind RT estimate. Here, we describe a method for estimating RT without prior knowledge of sound sources or room geometry. The diffusive tail of reverberation was modeled as an exponentially damped Gaussian white noise process. The time constant of the decay, which provided a measure of the RT, was estimated using a maximum-likelihood procedure. The estimates were obtained continuously, and an order-statistics filter was used to extract the most likely RT from the accumulated estimates. The procedure was illustrated for connected speech. Results obtained for simulated and real room data are in good agreement with the real RT values.

I. INTRODUCTION

The estimation of room reverberation time (RT) has existed in the field of engineering and architectural acoustics for nearly a century. The RT of a room specifies the duration for which a sound persists after it has been switched off. The persistence of sound is due to the multiple reflections of sound from the various surfaces within the room. RT is commonly referred to as the T_{60} time since the accepted measure of persistence of sound is the time to decay 60 dB below its peak value. Since reverberation results in temporal and spectral smearing of the sound pattern, thus distorting both envelope and fine structure, the RT of a room provides a measure of the listening quality of the room. This is of particular importance in speech perception where it has been noted that speech intelligibility reduces as the RT increases, due to masking within and across phonemes ([1, 5, 11]). State-of-the-art hearing aids, or other audio processing instruments, which implement signal processing strategies tailored to specific listening environments, are expected to have the ability to evaluate the characteristics of the environment, and turn on the most appropriate strategy. Thus, a method that can characterize the RT of a room from passively received microphone signals represents an important enabling technology.

In the early 20th century, Sabine ([17]) provided an empirical formula for the explicit determination of RT based solely on the geometry of the environment (volume and surface area) and the absorptive characteristics of its surfaces. Since then, Sabine's reverberation time equation has been extensively modified and its accuracy improved (see [7] for a historical review of the modifications), so that today, it finds use in a number of commercial software packages for the acoustic design of interiors. Formulae for calculation of RT are used primarily for the design of concert halls, classrooms, and other acoustic spaces where the quality of the received sound is of greatest importance, and the extent of reverberation has to be controlled. However, in order to determine the RT of existing environments, both the geometry and the absorptive characteristics have to be first determined. When these cannot be easily determined, it is necessary to search for other methods, such as those based purely on controlled recordings of sounds in the environment to be tested.

Schroeder ([18, 19]) presented a method for estimating RT based solely on the recording of an acoustic signal, radiated into the test enclosure. The method obviates some of the problems occurring in situations where the geometry and surface absorption characteristics are unknown. A burst of broad- or narrow-band noise is radiated into the test enclosure until it reaches steady state, and it is then abruptly switched off. The method tracks the sound energy decay following sound cessation ([18]). This method, referred to as Schroeder's backward integration method, while theoretically and practically important, has some limitations. Specifically, the sound used for measuring RT must be stationary and uncorrelated, and the precise time of sound offset must be known.

While Schroeder's method has been improved over the years (see [2, 22] for example), the improvements do not lift the restrictions placed on the applicability of Schroeder's method. At present, a "blind" method that requires no knowledge of the geometry, absorptive characteristics, and sound source characteristics or offset time of the sound, is unavailable. Partially blind methods have been developed where the characteristics of the room are "learned" using neural network approaches ([3, 12, 21]), or where some form of segmentation procedure is used for detecting gaps in sounds so that the sound decay curve can be tracked ([8]). The only other method that may be described as truly blind is "blind dereverberation", where sound source recovery is attempted by deconvolving the room output with the unknown room impulse response. In principle, this method can be used for extracting the RT, but there are serious drawbacks that limit its applicability. Namely, the room impulse response must be minimum phase, a condition that is rarely satisfied ([10, 13]).

Here we attempt to address several of the drawbacks found in existing methods by providing a blind approach that requires only one recording microphone. It does not require that a test signal be radiated into the test enclosure (as in Schroeder's method) or that the geometry and absorption characteristics of the test enclosure be known (as with the Sabine type formulae). The system performs blind estimation based on a decay curve model describing the reverberation characteristics of the enclosure. Sounds in the test enclosure (speech, music, or other pre-existing sounds) are continuously processed and a running estimate of the reverberation time is produced by the system using a maximum-likelihood parameter estimation procedure. A decision-making step then collects estimates of RT over a period of time and arrives at the most likely RT using an order-statistics filter.

II. THEORY

A model for blind estimation of reverberation time is presented, followed by an algorithm for implementation, and a decision-making strategy for selecting the estimate best representing the reverberation time of listening rooms.

Before describing the model, we motivate the work with an example. The recorded response of a room to an impulsive sound source (a hand-clap) is shown in Fig. 1A. As can be expected, there are strong early reflections followed by a decaying reverberant tail. If the early reflections are ignored, the decay rate of the tail can be estimated from the envelope. A common and widely used measure of the decay time is the T_{60} time first defined by Sabine [17], which measures the time taken for a sound to drop 60 dB below the level at sound cessation. In practice, a decaying sound in a real environment reaches the ambient noise floor, thus limiting the dynamic range of the measured sound to values less than 60 dB, and so it is not possible to directly measure T_{60}. Instead, the time to reach −25 dB or −35 dB from a reference level of −5dB is often used [18]. These values can be extrapolated to obtain T_{60}. Figure 1C shows the measurement of T_{60} from the hand-clap data using Schroeder's method [18] described below. Schroeder's method, while exact, suffers from the drawback that the precise instant of cessation of sound must be known, and there must be a sufficiently long period of silence to perform the estimation. Thus, it is not amenable to online implementation when sounds such as connected speech are present.

We begin with a model for the diffusive or reverberant tail of sounds in a room. This refers to the dense reflections that follow the early reflections. All that can be said about them is that they are the result of multiple reflections, and appear in random order, with successive reflections being damped to a greater degree if they occur later in time. Traditionally, and dating back to Sabine, the decay envelope has been modeled as an exponential with a single time-constant. Since the dense reflections are assumed to be uncorrelated, a convenient and highly simplified model is to consider the reverberant tail to be an exponentially damped uncorrelated noise sequence with gaussian characteristics. The model does not include the direct sound or early reflections. The goal is to estimate the time-constant of the envelope.

A. Model of sound decay

We assume that the reverberant tail of a decaying sound y is the product of a fine structure x that is random process, and an envelope a that is deterministic. A central assumption is that x is a wide-band process subject to rapid fluctuations, whereas the variations in a are over much longer time-scales. Here, we will provide a statistical description of the reverberant tail with the goal of estimating the decay time-constant of the envelope.

Let the fine structure of the reverberant tail be denoted by a random sequence $x(n)$, $n \geq 0$ of independent and identically random variables drawn from the normal distribution $\mathcal{N}(0, \sigma)$. Further, for each n we define a deterministic constant $a(n) > 0$. The model for room decay then suggests that the observations y are specified by the sequence $y(n) = a(n)x(n)$. Due to the time varying term $a(n)$, the $y(n)$ are independent but not identically distributed, and their probability density function is $\mathcal{N}(0, \sigma a(n))$. That is, the constant $a(n)$ modulates the instantaneous power of the fine structure. For purposes of estimating the room decay time, we consider a finite sequence of observations, $n = 0, \ldots, N$ where N will be referred to as the estimation interval or estimation window length. For notational simplicity, denote the N-dimensional vectors of y and a by $\boldsymbol{y}$ and $\boldsymbol{a}$, respectively. Then

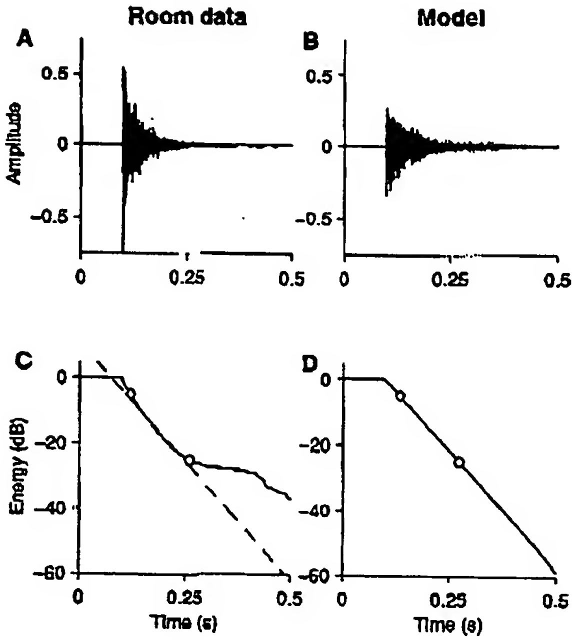

Figure 1: Temporal decay of a hand–clap at $t = 0.1$ s as recorded by a microphonc (left column) and the model matching the reverberation (right column). (A) The recorded sound shows strong early reflections followed by a reverberant tail. Direct sound is excluded from the trace. (B) Model matching the reverberant tail shown in (A). Direct and early reflections are excluded. The model is a Gaussian white noise process damped by a decaying exponential, parametrized by the noise power σ and decay time-constant τ. (C) Decay time-constant estimated from Schroeder's backward integration method [18] between −5 dB (◇) and −25 dB (o). Slope of linear fit (dashed line) yields $\tau = 59$ ms ($T_{60} = 0.4$ s). (D) Decay curve for model has identical slope everywhere following sound offset, and captures the most significant part of decay (−5 dB to −25 dB).

the likelihood function of y (the joint probability density), parameterized by a and σ is

$$L(y;\, a,\, \sigma) = \frac{1}{a(0)\ldots a(N-1)}(\frac{1}{2\pi\sigma^2})^{N/2}\exp(-\frac{\sum_{n=0}^{N-1}(y(n)/a(n))^2}{2\sigma^2}), \tag{1}$$

where a and σ are the $(N+1)$ unknown paramaters to be estimated from the observation y. The likelihood function given by (1) is somewhat general, and while it is possible to develop a procedure for estimating all $(N+1)$ parameters, suitable simplifications can be made when modeling sound decay in a room. Let a single time-constant τ describe the damping of sound envelope during free decay. Then the sequence $a(n)$ is uniquely determined by

$$a(n) = \exp(-n/\tau). \qquad (2)$$

Thus, the N-dimensional parameter $\mathbf{a}$ can be replaced by a scalar parameter a that is expressible in terms of τ and a single parameter $a = \exp(-1/\tau)$, so that

$$a(n) = a^n. \qquad (3)$$

Introducing (3) into (1) yields

$$L(y;\, a,\, \sigma) = (\frac{1}{2\pi a^{(N-1)}\sigma^2})^{N/2} \exp(-\frac{\sum_{n=0}^{N-1} a^{-2n}\, y(n)^2}{2\sigma^2}). \qquad (4)$$

For a fixed observation window N, and a sequence of observations $y(n)$, the likelihood function is parameterized solely by the time-constant a and the diffusive power σ.

The model is shown in Fig. 1B with parameters a and σ matched to the experimental hand-clap data shown in Fig. 1A. Note that the model does not include the early reflections shown in panel A. The Schroeder decay curve for the model is shown in Fig. 1D with a T_{60} time of 0.4 s in agreement with the measured T_{60}. The agreement between model and real T_{60} time motivates the search for an algorithm that can optimally estimate the two parameters.

B. Maximum-likelihood estimation

Given the likelihood function, the parameters a and σ can be estimated using a maximum-likelihood approach [15]. First, we take the logarithm of (4) to obtain the log-likelihood function

$$\ln L(y;\, a,\, \sigma) = -\frac{N(N-1)}{2}\ln(a) - \frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1} a^{-2n}\, y(n)^2. \qquad (5)$$

Then, we differentiate the log-likelihood function (5) with respect to a to obtain the score function [15]

$$\begin{aligned} s_a(a;\, y,\, \sigma) &= \frac{\partial \ln L(y;\, a,\, \sigma)}{\partial a}, & (6)\\ &= -\frac{N(N-1)}{2a} + \frac{1}{a\sigma^2}\sum_{n=0}^{N-1} n\, a^{-2n}\, y(n)^2. & (7)\end{aligned}$$

The log-likelihood function achieves an extremum when $\partial \ln L(y;\, a,\, \sigma)/\partial a = 0$. That is, when

$$-\frac{N(N-1)}{2a} + \frac{1}{a\sigma^2}\sum_{n=0}^{N-1} n\, a^{-2n}\, y(n)^2 = 0. \qquad (8)$$

The zero of the score function provides a best estimate in the sense that $\mathrm{E}[s_a] = 0$.

Denote the zero of the score function s_a, and satisfying (8), by a^*. It can be shown that the second derivative $\partial^2 \ln L(y;\, a^*,\, \sigma)/\partial a^2 < 0$, i.e., the estimate a^* maximizes the log-likelihood function.

The diffusive power of the reverberant tail or variance σ^2 can be estimated in a similar manner. Differentiating the log-likelihood function (5) with respect to σ, we have

$$\begin{aligned} s_\sigma(\sigma;\, y,\, a) &= \frac{\partial \ln L(y;\, a,\, \sigma)}{\partial \sigma}, & (9)\\ &= -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{n=0}^{N-1} a^{-2n} y(n)^2, & (10)\end{aligned}$$

which achieves an extremum when

$$\sigma^2 = \frac{1}{N}\sum_{n=0}^{N-1} a^{-2n} y(n)^2. \tag{11}$$

As before, it can be shown that the $\mathrm{E}[s_\sigma] = 0$. Denote the zero of the score function s_σ, and satisfying (11), by σ^*. It can be shown that the second derivative $\partial^2 \ln L(y;\, a,\, \sigma^*)/\partial\sigma^2 < 0$, i.e., the estimate σ^* maximizes the log–likelihood function. Note that the maximum-likelihood equation given by (8) is a transcendental equation and cannot be inverted to solve directly for a^*, whereas the solution of (11) for σ^* is direct. This is considered in detail later when an algorithm for estimation is developed.

Bounds on the estimate of a and σ are obtained from the variance of the score function, also called the Fisher information J. This is more conveniently expressed in terms of the derivatives of the score functions [15]. Given the parameter $\theta^{\mathrm{T}} = [a\ \sigma]$ and the score function $s_\theta^{\mathrm{T}}(y;\, \theta) = [s_a(y;\, a,\, \sigma)\ s_\sigma(y;\, a,\, \sigma)]$, we have

$$J(\theta) = -\mathrm{E}[\frac{\partial\, s_\theta^{\mathrm{T}}(y;\, \theta)}{\partial\theta}]. \tag{12}$$

From (7), (9), and (12), we have

$$J(\theta) = \begin{pmatrix} \frac{N(N-1)(2N-1)}{3a^2} & \frac{N(N-1)}{a\sigma} \\ \frac{N(N-1}{a\sigma} & \frac{2N}{\sigma^2} \end{pmatrix}. \tag{13}$$

By the Cramer–Rao theorem [15], a lower bound on the variance of any unbiased estimator is simply $J^{-1}(\theta)$, which is

$$J^{-1}(\theta) = \begin{pmatrix} \frac{6a^2}{N(N^2-1)} & -\frac{3a\sigma}{N(N+1)} \\ -\frac{3a\sigma}{N(N+1)} & \frac{\sigma^2(2N-1)}{N(N+1)} \end{pmatrix}. \tag{14}$$

From the asymptotic properties of maximum-likelihood estimators [15], we know that the estimates of a and σ are asymptotically unbiased and their variances achieve the Cramer-Rao lower bound (i.e., they are efficient estimates). Thus, if a^* and σ^* are the estimates obtained from the solutions of (8) and (11), the variance of the estimates are

$$\mathbf{E}[(a^* - a)^2] \geq \frac{6a^2}{N(N^2 - 1)}, \tag{15}$$

$$\mathbf{E}[(\sigma^* - \sigma)^2] \geq \frac{\sigma^2(2N - 1)}{N(N + 1)}, \tag{16}$$

with equality being achieved in the limit of large N. Since the variance of a and σ are $O(N^{-3})$ and $O(N^{-1})$, the estimation error can be made arbitrarily small if observation windows are made sufficiently large.

C. Algorithm for estimating decay time

Given an estimation window length and the sequence of observations $y(n)$ in the window, the zero of the score function (8) provides an estimate of a. The function is a transcendental equation that must be solved numerically using an iterative procedure. Since the estimate of σ can be obtained directly from (11), a two-step procedure was followed. First, an approximate solution for a^* from (8) was obtained, and next, the value of σ^* was updated from (11). The procedure was repeated, providing succesively better approximations to a^* and σ^*, and so converging on the root of (8).

Here we address the strategy for extracting the root in the smallest number of iterative steps. To gain an understanding of the root-solving procedure, we consider the example shown in Fig. 2. The function $a = \exp(-1/\tau)$ maps the room time-constant τ one-to-one and onto a as shown in Fig. 2A. For instance, consider a room time-constant of 0.1 s and a sampling rate of 16 kHz. Then the time-constant is 1600 samples, and so $a = 0.9994$ (filled circle). The signficance of the number becomes clear if we consider that when the time-constant is 0.03 s, then $a = 0.9979$, whereas for $\tau = \infty$, $a = 1$. Hence the geometric ratio is highly compressive and values of a for real environments are likely to be close to 1. Thus, the advantage of estimating a rather than τ is due to the bounded nature of

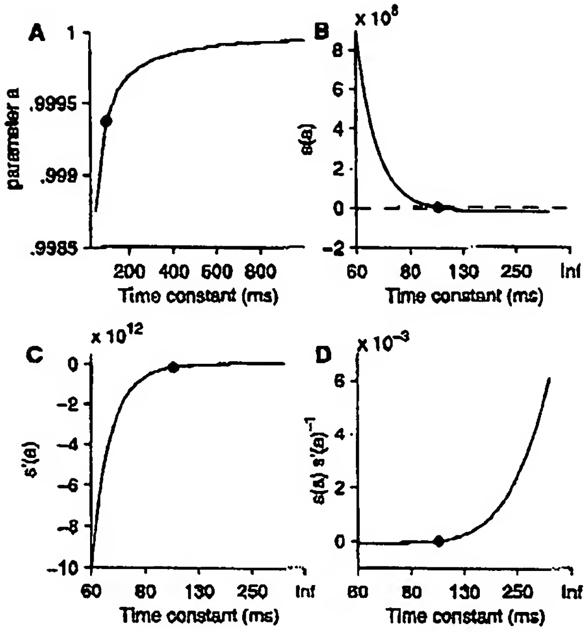

Figure 2: maximum-likelihood estimation (MLE) of room decay time-constant. (A) The time-constant of the exponential decay (τ, abscissa) is mapped to a parameter $a = \exp(-1/\tau)$ (ordinate) where τ is given in sampling periods. The function is monotone but highly compressive and maps $\tau \in [0, \infty)$ onto $a \in [0, 1)$. Filled circle shows $\tau = 100$ ms ($a = 0.9994$). (B) Score function (derivative of log likelihood function) $s_a(a)$ (ordinate), decreases rapidly as a function of a (abscissa, marked in time constants using the map in (A)). MLE of a is given by the root of $s(a)$ (filled circle). (C) The derivative $s_a\prime(a)$ as a function of a. At the root of s_a (filled circle), the derivative is negative. Note the nearly 8–12 orders of magnitude change in s_a and $s_a\prime$ for commonly encountered values of τ. (D) The ratio $s_a(a)/s_a\prime(a)$ (ordinate) as a function of a is the incremental step size of the Newton-Raphson procedure for finding the root of (8). It provides an estimate of the convergence properties of the root-finding algorithm. Sampling frequency was 16 kHz, and the log-likelihood function was calculated assuming a 400 ms window.

a. The score function s_a from (7) on the other hand, has a wide range (about 8 orders of magnitude, see Fig. 2B) and is zero at the room time-constant (filled circle). The gradient of the score function ds_a/da shown in Fig. 2C also demonstrates a wide range, but takes a negative value at the zero of s_a.

Thus, if we start with an initial value of $a_0^* < a$, the root-solving strategy must descend the gradient sufficiently rapidly. The standard method for solving this kind of nonlinear equation, where an explicit form for the gradient is available, is the Newton-Raphson method which offers second-order convergence [16]. The order of convergence can be assessed from $s_a\,(ds_a/da)^{-1}$ which is the incremental step size Δa in the iterative procedure (Fig. 2D). For example, with true value of $\tau = 100$ ms, Δa at intermediate values in the iteration can be as small as 10^{-6} when $a = 0.9993$ ($\tau = 90$ ms) or $a = 0.9995$ ($\tau = 120$ ms). This corresponds to an incremental improvement of about 0.01 ms every iteration, thus providing slow convergence if the initial value is far from the zero. On the other hand, the bisection method [16] guarantees rapid gradient descent but works poorly in regions where the gradient changes relatively slowly (such as near the true value of a). Furthermore, it guarantees only first-order convergence.

However, the specific structure of the root-solving problem can be exploited since the behavior of s_a is known. Here, both methods were used to obtain convergence to the root. First, the root was bisected until the zero was bracketed as close as possible, after which the Newton-Raphson method was applied to polish the root. For the example shown, the root bracketing was accomplished in about 8 steps and the root polishing in 2-4 steps. In contrast, with the same initial conditions, the Newton-Raphson method took about 500 steps to converge. Taken together, the analysis presented here suggests that the estimation procedure is feasible and does not lead to significant errors even though values of a for real rooms are close to 1, and the score function and its derivative vary over many orders of magnitude. While other algorithms are possible, these are not dealt with here.

D. Strategy for assigning the correct decay time from the estimates

The theory presented in the preceding section provides one estimate of a and σ in a given time frame of N samples. By advancing the frame as the signal evolves in time, a series of estimates a_k^* are obtained, where k is the time frame. We assume that after a sufficient number of estimates have been obtained, a decision is made regarding the true value of the decay time-constant. In this section, we present a strategy for making such a decision.

Since we are considering blind estimation, the input is unknown, and so the model will fail when: 1) an estimate is obtained in a frame that is not ocurring during a free decay. This includes regions where there is sound onset or sound is ongoing. In these periods, the MLE scheme can provide widely fluctuating or implausible estimates due to model failure. 2) During a region of free decay initiated by a sound with a gradual rather than rapid offset. In this case, the offset decay of the sound will be convolved with the room response, prolonging the sound even further and so, the estimated time-constant will be larger than the real room time-constant. Gradual offsets occur in many natural sounds, such as terminating vowels in speech. We address both issues here and provide a strategy for selecting the correct room time constant.

In the first case where the estimation frames do not fall within a region of free decay, many of the time frames will provide estimates of a close to unity (i.e., infinite τ), or implausible values. On the other hand, the estimates will accurately track the true value when a free decay occurs. Intuitively, a strategy for selecting a from the sequence a_k^* is guided by the following observation: the damping of sound in a room cannot occur at a rate faster than the free decay, and thus all estimates a^* must attain the true value of a as a lower bound. The bound is achieved only when a sound terminates abruptly, upon which the model conditions will be satisfied, and the estimator will track the true value of the time-constant.

While it seems intuitive to set $a = \min\{a_k^*\}$, it should be recognized that even during a free decay the estimate is inherently variable, and so selecting the minimum is likely to underestimate a.

A robust strategy would be to select a threshold value of a^* such that the left tail of the probability density function of a^*, $p(a^*)$, occupies a pre-specified percentile value γ. This can be implemented using an order statistics filter specified by

$$a = \arg\{P(x) = \gamma : P(x) = \int_0^x p(a^*)\, da^*\}. \tag{17}$$

For a unimodal symmetric distribution with $\gamma = 0.5$ the filter will track the peak (median) value. Order statistics filters play an important role in robust estimation, especially when data is contaminated with outliers [14], as is the case here. It should be noted that for γ values approaching 0 the filter (17) performs like the minimum filter $a = \min\{a_k^*\}$ suggested above.

In the second case described above, where the sound offset is gradual, $p(a^*)$ is likely to be multimodal since sound offsets (such as terminating phonemes in speech) will have varying rates of decay, and their presence will give rise to a multiple peaks. The strategy then, is to select the first dominant peak in $p(a^*)$ when a^* is increasing from zero (i.e., leftmost peak). That is,

$$a = \min \arg \{dp(a^*)/da^* = 0\}, \tag{18}$$

where the minimum is taken over all zeros of the equation. If the histogram is unimodal but asymmetric, the filter tracks the mode and resembles the order-statistics filter.

In conversational speech, where peaks cannot be clearly discriminated or the distribution is multimodal, (17) can be employed by choosing a value of γ based on the statistics of gap durations. For instance, if gaps constitute approximately 10% of total duration, then $\gamma = 0.1$ would be a reasonable choice. A judicious choice of γ can result in the filter performing like an edge detector, since it captures the transition from larger to smaller values of the time-evolving sequence a_k^*.

The decision strategies, (17) and (18), were used to validate the model in simulated and real environments (see Results).

III. EXPERIMENTAL METHODS

In addition to simulations, the MLE approach was validated with real room data. The experimental methods and data analysis procedures are described in the following sections.

A. Sound recordings

To validate the MLE method, sounds recordings were made in several rooms, corridors and an auditorium, with the aim of determining their reverberation times. Sound stimuli that were used included 18-tap maximum length (ML) sequences (period length of $2^{18} - 1$), clicks (100 μ s), hand-claps, word utterances, and connected speech from the TIMIT corpus. Recordings were made using a single Sennheiser MK-II omni-directional microphone (frequency response 100-20000 Hz). Microphone cables (Sennheiser KA 100 S-60) were connected to the XLR input of a portable PC-based sound recording device (Sound Devices USBPre 1.5). The recorder transmitted data sampled at 44.1 kHz to a laptop computer (Compaq Presario 1700, running Microsoft Windows XP) via a USB link. Sound stimuli, stored as single-channel pre-sampled (44.1 kHz) WAV files, were played through the headphone output of the laptop, amplified by a power amplifier (ADCOM GFA-535II) and presented through a loudspeaker (Analog and Digital Systems Inc., ADS L200e, 85-20000 Hz). Data acquisition and test material playback were controlled by a custom written script in MATLAB (The MathWorks Inc.,) using the Sound PC Toolbox (Torsten Marquardt).

B. Measurement of T_{60} time using Schroeder's method

To validate the estimation procedure developed in this work, experimentally recorded data from real listening environments were processed using the ML procedure and compared to results obtained from a widely used method developed by Schroeder [18]. The method proposed by Schroeder determines the decay time-constant from the sound decay curve following the cessation of a broad- or narrow-band noise burst. Briefly, if $r(t)$ is the measured decay curve from a single trial, then the mean squared average of the decay curve $s(t)$ over a large number of trials is related to $r(t)$ by

$$\mathrm{E}[s^2(t)] = \int_t^\infty r^2(x)\,dx, \tag{19}$$

where the sound is assumed to switch off at $t = 0$. Schroeder's method, called the backward integration method, can be applied to a single broad-band channel or to multiple narrow-band channels. The recorded data were filtered offline in ISO one-third octave bands (21 bands with center frequencies ranging from 100-10000 Hz) using a fourth-order Type II Chebyshev band-pass filter with stopband ripple 20 dB down. The output from each channel was processed by the ML procedure and Schroeder's method using (19). For the broad-band estimation, the microphone output was processed directly using the two methods.

Due to the limited dynamic range of sounds in real environments, Schroeder's method requires the specification of a decay range. The decay ranges normally used are from −5 dB to −25 dB (20 dB range), and from −5 dB to −35 dB (30 dB range). The decay curves in each range were fitted to a regression line using a nonlinear least squares fitting function (function nonlinsq provided by MATLAB). The fitted function was of the form $A\,a_d^n$, where A is a constant, n is sample number within the decay window, and a_d is the geometric ratio related to the decay time-constant by $a_d = \exp(-1/\tau_d)$. This is in contrast to the model (2) which assumes an exponentially decaying envelope with time-constant τ, whereas the decay curve is obtained by squaring the signal. Hence, $\tau_d = \tau/2$. Since decay curves were fitted to the −5 to −25 dB, and −5 to −35 dB drop-offs, two estimates of the time-constant were obtained. The fitted line was extrapolated to obtain the T_{60} time (in seconds) for each of the estimates using

$$T_{60} = \frac{6}{\log_{10}(e^{-1})\,\log_e(a_d)} = \frac{-6\,\tau_d}{\log_{10}(e^{-1})} = 13.82\,\tau_d. \tag{20}$$

The same procedure was followed for determining the time-constant from the broad-band signal. It should be noted that the ML procedure does not requirethe specification of a decay range, but only the specification of the estimation window length. So only one estimate per band is obtained.

C. Verification of MLE procedure with ideal stimuli

Microphone data were processed using the MLE procedure to obtain a running estimate of the decay time-constant. For model verification, estimation was performed on: 1) the segment following the cessation of a maximum-length sequence or a hand-clap, and 2) the entire run of a string of isolated word utterances. These were considered ideal stimuli since they fulfilled the model assumptions of free decay or long gaps between sounds. The estimates were binned for each run and a histogram was produced. The histogram was examined for peaks, and the time-constant was selected using the order-statistics filter (18) if there were multiple peaks, or (17) if the histogram was unimodal. The estimate $\hat{a}$ so obtained was used to calculate the T_{60} time (in seconds) using the formula

$$T_{60} = \frac{3}{\log_{10}(e^{-1})\,\log_{e}(\hat{a})} = \frac{-3\,\tau}{\log_{10}(e^{-1})} \approx 6.91\,\tau. \tag{21}$$

In theory, the T_{60} expressions given by (20) and (21) are identical due to the relationship between τ and τ_d. However, the calculated values may differ, and this can be ascribed to either model inadequacies or discrepancies in measurement and analysis.

D. Verification of MLE procedure for speech

The performance of the MLE was also verified using connected speech played back in a real room. Test material were connected sentences from the Connected Speech Test (CST) corpus. Estimates from non-overlapping 1 s intervals were binned to yield a histogram, and the first dominant peak from the left of the histogram was selected to determine the room time-constant. The procedure for calculating T_{60} time followed (21).

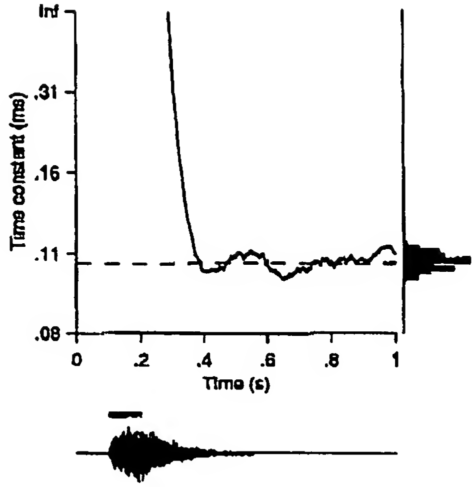

Figure 3: Illustration of procedure for continuous estimation of room decay time. A burst of white noise was applied at time $t = 0.1$ s (black bar, bottom trace, 100 ms duration). Simulated room output (bottom trace) shows the build-up and decay of sound in the room. A running estimate of the parameter a in 200 ms windows is shown in the graph (ordinate, a shown in units of time-constant). True value of room decay time (100 ms) is shown as horizontal dashed line. The estimation window was advanced by one sample to obtain the trace, with each point at time t being the estimate in the window $(t - 0.2, t]$. During the build-up and ongoing phase of the sound, estimated a sometimes exceeded 1 (i.e., negative values of τ). These were discarded and are not shown. As the window moved into the region of sound decay ($t > 0.3$ s), the estimates converged to the correct value. A histogram of the estimated time-constant is shown to the right of the trace. An order-statistics filter, such as the mode of the histogram, can be used to extract the room time-constant. Sampling rate was 16 kHz.

IV. RESULTS

The estimation procedure was applied to a variety of data sets, including simulated data and real room responses. To illustrate the methods and identify the strengths and deficiencies of the estimation procedure, we first consider simulated data sets. Subsequently we will provide results for real data that validate the room time-constant estimates, and compare these to results from Schroeder's method.

A. Broad band white noise bursts in simulated rooms

A 100 ms burst of broad-band white noise (8 kHz bandwidth) was radiated into a simulated room having a decay time-constant $\tau = 100$ ms (Fig. 3). Room output shown in the bottom trace of Fig. 3 shows the characteristic rise and decay of sound folowing onset and offset of noise burst (horizontal bar). The graph shows the running estimate of decay time-constant obtained in a 200 ms time window by advancing every sample. Since time frames up to about $t = 0.3$ s are not regions of free decay, the estimator tended to produce values of $a > 1$. When this was observed in the root-bracketing step of the estimate, the root-solving procedure was aborted. Thus all estimates of a were bounded above by 1. It can be seen that when the window crosses into the region of free decay, the estimator output stabilizes at the true value (horizontal dashed line). A histogram of the time-constant estimates (right axis) was input to the order statistics filter (17) with $\gamma = 0.5$. The reported time-constant from the filter was $a = 101$ ms.

For comparison, the procedure was repeated with the simulated noise burst to mimic anechoic conditions. The histogram of a^{*} demonstrated a strong peak at $a = 1$ ($\tau = \infty$) (not shown). This showed that no consistent estimate of a was found. While this result acted as a control, histograms showing strong peaks at $a = 1$ are to be expected in anechoic or open spaces.

B. Effect of filter length on estimation

A parameter that is critical for estimation performance is the window length N specified in (8). Small window lengths are expected to increase the variance of the estimate, as also indicated by the Cramer-Rao lower bound (15). This is shown in Fig. 4. A burst of white noise (100 ms duration) was convolved with a simulated room impulse response ($\tau = 100$ ms) and the estimator tracked the decay curve. The left column depicts the running estimate for four different window lengths: 0.5τ, τ, 2τ, and 4τ, top to bottom, respectively (dashed line: τ). The right column depicts the corresponding histograms. As

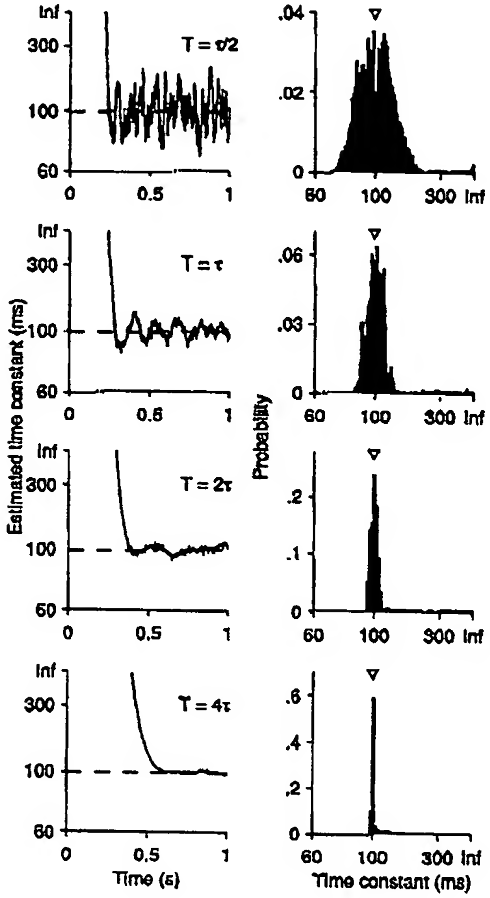

Figure 4: Effect of estimation window length on the variance of the estimate. The simulation shown in Fig. 3 was repeated for windows of duration 0.5τ, τ, 2τ, and 4τ (top to bottom), where $\tau = 100$ ms is the true value of the room time-constant. Left column shows the running estimate of parameter a (ordinate, shown as time-constant in ms) as a function of time (abscissa). The right column shows the histogram of the estimates. The variance of the estimate decreases with increasing window length; however there is no bias introduced by the estimator as evident from the location of the peak in the histogram (arrowheads mark true value of τ).

window length increased, the MLE procedure gave improved estimates. We concluded that increasing window length reduced the variability in the estimates, and did not introduce significant bias.

While it is desirable to have a large filter length, in practice this is limited by the duration and occurence of gaps between sound segments in the room output. Ideally the filter length should be of the order of τ or longer, but if the gaps are short, then increasing the filter length beyond the mean gap will produce undesirable end effects where the next sound segment creeps into the window. Thus, the window length should not be less than one-half or one-third of τ, but the upper limit is dictated by the mean duration of gaps.

C. Speech sounds in simulated room

The examples considered above illustrated the performance of the algorithm when the input was broad band white noise. To be applicable in realistic conditions, the algorithm must perform in a variety of conditions and signal types, most notably in those that include speech. Speech represents an example where the algorithm is expected to perform poorly, since it is nonstationary and nongaussian. Further, the offset transients in speech sounds (even for plosives) have a decay time that can vary from 5-40 ms. Thus, estimation of decay times with speech presents a particular challenge to the algorithm. We took a sequence of 15 distinct and isolated words recorded in an anechoic environment at a sampling rate of 20 kHz. These included 11 consonant-vowel-consonant words (/p,b,g/V/d/, e.g., "bed"), and 4 consonant-vowel words (/b/V/, e.g., "bay") separated by a mean interval of 200 ms. These were convolved with a simulated room impulse response having time-constant $\tau = 100$ ms. The task of the estimator was to track the decays for the entire duration of the sequence (approximately 11.4 s). The control condition was the clean input (i.e., anechoic). The results are shown in Fig. 5. Four different filter lengths were used as in Fig. 4. For the control condition (left column) no reliable estimates were produced for the smallest three windows (top three panels) since the histogram peaked at values of τ approaching ∞. For the simulated room response (right column), the peak shifted towards the true value of τ, with the best estimates being obtained for the largest window size of 4τ (right column, bottom row). In all the histograms the peak was located at about 115 ms (arrow). This estimate deviated from the real time-constant of 100 ms due to the lack of sharp transients in the clean speech. A gradual sound offset tends to

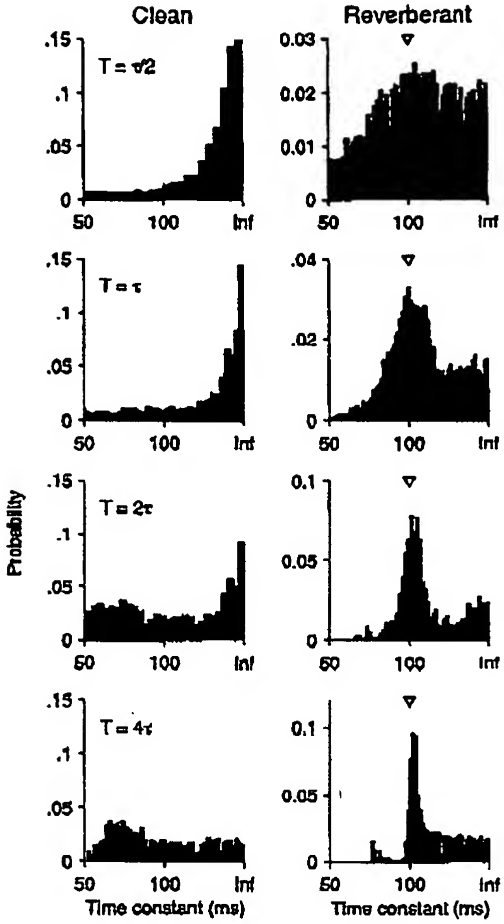

Figure 5: Estimation of room time-constant from speech. Fifteen words recorded in an anechoic (clean) environment (200 ms inter-word spacing) were convolved with a room model (τ = 100 ms). Histograms of decay time-constants were estimated from clean (left column) and simulated reverberant responses (right column), and are shown for window durations 0.5τ, τ, 2τ, and 4τ (top to bottom). The histogram for clean speech served as a control. Description follows Fig. 4. Estimation from reverberant speech produces a clearly defined peak, especially for the longer window lengths, albeit with a small bias (right column, 2τ and 4τ). The bias can be attributed to the gradually decaying offsets inherent in speech so that the resultant decay is a convolution of speech offset and the room response. For the control condition (left column), the offset decay is visible only in the bottom two rows where the histogram exhibits a broad bump between 50 and 100 ms. The fifteen words included 11 /p,b,g/V/d/ and 4 /b/V/ sampled at 20 kHz.

prolong the reverberated sound even further. This can be seen in the "anechoic" control condition where a small peak is noticeable when window size is 4τ (bottom panel, left column). The peak occurs around 60 ms, and corresponds to the gradual offsets of speech sounds. Thus, this introduces a bias in the estimates under reverberant conditions.

The results of the preceding sections demonstrate the importance of selecting a suitable estimation window length. The choice of window length determines the variability of the estimates, and is critical to obtaining a histogram with a clearly resolved peak at the true value of the room time-constant. Since the decision-making scheme uses an order statistics filter, which is a nonlinear operation, the robustness to the histogram variability are not known. Bimodal and multimodal histograms may be obtained if there is fluctuating background noise or the sound segments have an intrinsic offset decay rate (as shown above).

D. The effect on estimation of offset decay in speech

The preceding section introduced the problem of estimating room decay time-constant when the input signal exhibited varying offset decays. Here we examine in greater detail the performance of the estimator with input comprising a single word (/b/V/, "bough"). The word was recorded under anechoic conditions and presented to the estimator without modification so that the effect of the vowel offset could be determined. The results are shown in Fig. 6. The terminating vowel has a gradually decaying offset (top panel). Estimation of the offset decay was performed from $t = 0.45$ s (vertical dashed line). Two procedures were employed. First, the envelope was extracted from the analytic signal, windowed, and filtered to eliminate frequency components above 100 Hz. The envelope is shown in the middle panel (heavy outline). The envelope was then squared and transformed to a decibel scale, and the decay time-constant was estimated using a nonlinear least square fit in windows of duration 0.4 s (horizontal bar). Estimates were obtained by sliding the window forward in steps of one sample. Note that the time at which an estimate is reported for any given window is the end point of the window. The estimate

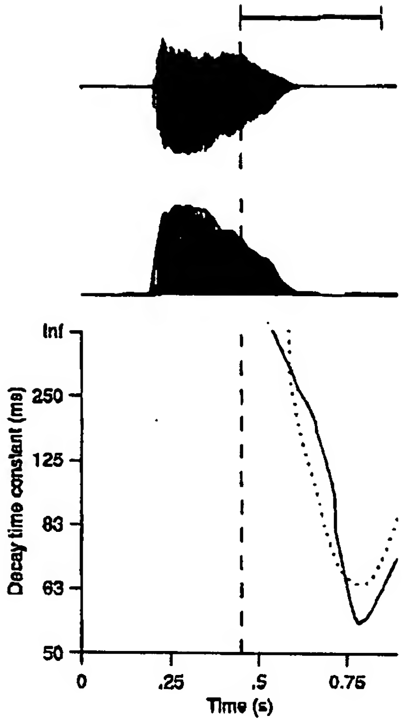

Figure 6: Illustration of decay time estimation when a terminating phoneme is encountered. The word "bough" recorded under anechoic (clean) conditions (top row) has a gradually decaying offset. The envelope was extracted by filtering the absolute value of the analytic signal (second row, heavy outline), and its decay rate was estimated for the segment following the dashed line using two methods (bottom row). Overlapping segments (duration given by bar, with step size indicated by the thickness of the vertical end) were converted to a decibel scale and the decay time-constant obtained by a least squares fit to a straight line (dotted trace). The same segments were analyzed using the MLE algorithm to obtain a second estimate of the decay time-constant (solid trace). While the estimators provide somewhat different results, they are in qualitative agreement. Both methods suggest that the fastest decay time-constant is in the range of 50-70 ms (see also Fig. 5). These results suggest that speech segments will introduce a bias when estimation is carried out in reverberant environments.

for the window indicated by the bar, for instance, is plotted at time $t = 0.85$ s. A curve of the estimated time-constants was thus obtained (dotted curve, bottom panel). The MLE procedure was applied to the same segments and produced an independent estimate of the the decay time-constant (solid line, bottom panel). While the estimates differ somewhat, they are in qualitative agreement. Both procedures indicate that the terminating vowel had a time-dependent decay rate, and the greatest rate was between 50 and 70 ms.

The results confirm the presence of the peak in Fig. 5 (left column, bottom panel), although the histogram shown in Fig. 5 was obtained for a sequence of 15 words. The analysis shown in Fig. 6 also indicates the reason for estimation bias under reverberant conditions using speech samples. Taken together, the results suggest that the factors responsible for estimation performance are: the presence of adequate numbers of gaps, sharp offset transients, and estimation window length.

E. Validation of method

The results demonstrate that estimation of decay time in room response is possible for a variety of sounds including impulses, noise bursts and speech. While we have shown that a reasonable agreement exists with a nonlinear least squares fit to the data (Fig. 6), a more careful evaluation is necessary to determine conditions under which the MLE procedure is likely to provide accurate estimates. Here we establish that the estimated decay times are comparable to decay times obtained from Schroeder's method [18]. Furthermore, any data collected must be under sufficiently realistic conditions where there is background noise and where the testing sound is not subject to experimental control. A comparison of MLE performance with the standard method in real environments will therefore establish the utility of the method.

We compared the estimates using the backward integration method proposed by Schroeder [18] in both single-channel (i.e., the broad-band signal), and multi-channel frameworks (i.e., narrow-band signals occupying ISO one-third octave bands). Since Schroeder's method requires a fitting procedure to estimate the time-constant, a decay range (either 20 or 30 dB below a reference level of −5 dB) was selected (see Section III). The MLE procedure does not require the specification of such a range. We first report on the comparison between the methods using a hand-clap in a small office (8x3 m). Subsequently we will summarize results obtained in a number of rooms of different sizes.

Figures 7A, B depict a hand-clap event and its spectrogram, respectively. The data in panel A is the same as shown in Fig. 1A, except that Fig. 7A also includes the direct sound. The noise level in the room was 50 dB SPL, and peak sound pressure level resulting from the hand-clap was 85 dB SPL. The decay curve obtained using Schroeder's backward integration method is shown in Fig. 7C, normalized so that peak SPL was 0 dB. This is the broad-band curve obtained by integrating the recorded microphone signal. A straight-line fit to the 20 dB drop-off point (circle) from a reference level of -5 dB (lozenge) yielded $\tau = 56$ ms ($T_{60} = 0.39$ s). The discrepancy between this value and that presented in Fig. 1 ($\tau = 59$ ms) was due to the inclusion of the direct sound in Fig. 7. The windows over which the 20 dB drop-off was computed were not identical for the two cases. The data were run through the MLE procedure and a histogram of estimates was obtained, and the decay time-constant was calculated from the peak of the histogram using (17). This gave an estimate $\tau = 53$ ms ($T_{60} = 0.37$ s), which is in good agreement with the estimate obtained from Schroeder's method. Note that the estimates reported in this work are based on a single realization of sound events. The normal practise is to average over large numbers of trials. Since our goal is to develop an online estimation procedure, we felt that it would be more realistic to use a single trial.

To test a range of room RTs, ISO one-third octave band analysis (exceeding 1 kHz center frequency) was performed in three environments. These were 1) the moderately reverberant room described above (Fig. 7), 2) a highly reverberant circular foyer, and 3) a highly reverberant enclosed cafeteria. In all cases, the signal was a hand-clap generated 2 m in front of the recording microphone, with peak energy exceeding 90 dB SPL. Output from the band-pass filters were analyzed using the MLE procedure, and the τ value for each band was obtained from the histogram by selecting the dominant peak. For Schroeder's method, a 20 dB decay range was used. Figure 8 shows the T_{60} estimates from Schroeders method (abscissa) versus the ML estimates (ordinate) for each ISO one-third band (open symbols), and the average over these bands (closed symbols).

The figure shows that the variability of estimates for highly reverberant environments increases with increasing mean RT for both methods. However, the two methods are in good agreeement especially in the high-frequency bands (the single outlier falling below

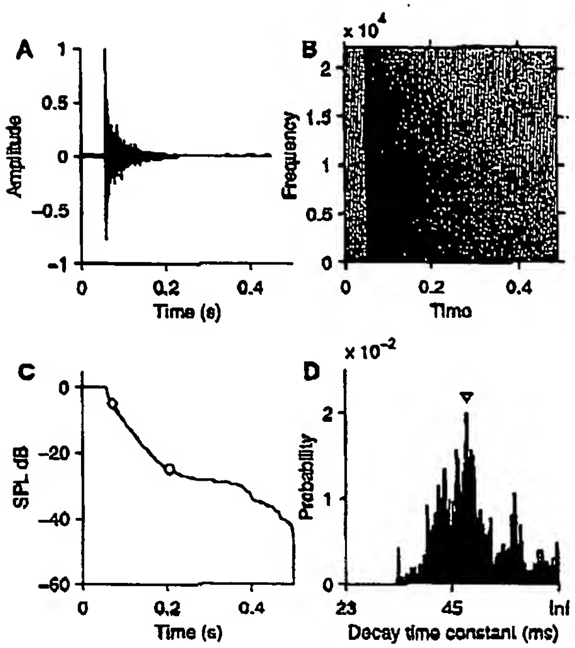

Figure 7: Estimation of decay time-constant from real room data. (A) The room response to a hand-clap (same as Fig. 1A but includes the direct sound). (B) Spectrogram of the hand-clap demonstrates a sharp broadband onset transient and the decay as a function of frequency. (C) The decay time-constant was estimated using Schroeder's backward impulse integration method in the -5 dB (lozenge) to -25 dB (circle) range, followed by a least squares fit to a straight line to obtain the time-constant ($\tau = 56$ ms, $T_{60} = 0.39$ s). (D) Histogram of decay times obtained from signal shown in A using MLE. The median value of the histogram (arrow, $\tau = 53$ ms, $T_{60} = 0.37$) is in good agreement with the estimate obtained using Schroeder's method.

the diagonal in Fig. 8 is the lowest center frequency used in the analysis, namely 1 kHz). The agreement between the methods is best when the T_{60} values are averaged over all bands (filled symbols), as is usually reported in the literature.

A more extensive test to determine the variability in estimates across different environments, and between bands, was performed in 12 environments, including small office rooms, an auditorium, large conference rooms, corridors, and building foyers. The data were analyzed as in Fig. 8 and are shown in Fig. 9A. In comparison with Schroeder's method, the MLE procedure consistently overestimated T_{60} in low to moderately reverberant environments ($T_{60} < 0.3$ s) whereas it underestimated the reverberation time for

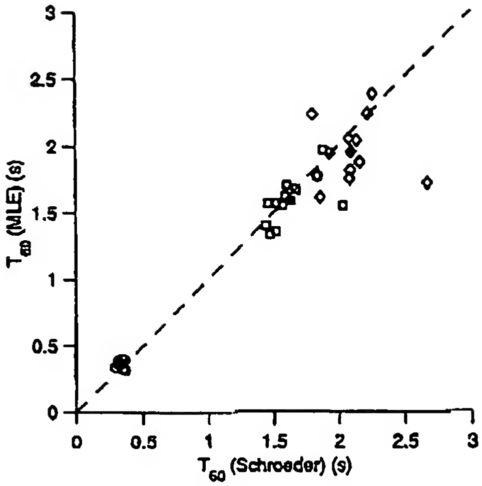

Figure 8: Comparison of Schroeder's method and the MLE procedure for T_{60} times obtained in one-third octave bands. Three environments were tested: a moderately reverberant environment (circles; the environment is the same as shown in Fig. 7), a highly reverberant circular foyer (squares), and a highly reverberant enclosed cafeteria (diamonds). In each environment, a single hand-clap was filtered using a bank of ISO one-third octave band-pass filters with center frequencies exceeding 1 kHz. Ordinate shows the best estimates obtained from the MLE procedure for each band. Abscissa shows the T_{60} times obtained from Schroeder's method Averages over all bands for each environment are shown as filled symbols. The diagonal dashed line is shown for reference, and points lying close to this line suggest good agreement between the two procedures. Agreement is best when the T_{60} values are averaged over all the bands.

more reverberant environments ($T_{60} > 1.3$ s). There was good agreement between the two methods for intermediate ranges. The average T_{60} over all bands (filled squares) were however, in good agreement. Broad-band estimates were made using the same procedures but without band-pass filtering of recorded signals. These are shown in Fig. 9B. The trend in the estimates was similar to that observed with narrow-band signals, except for one outlier. The outlier along with three other data points were obtained in a large auditorium. The latter three were obtained with a source-to-microphone distance of 1.5 m, whereas for the outlier the distance was 4 m. The sound levels were not adjusted to compensate for the distance, and hence the test corresponding to the outlier was at a lower SPL, resulting in reduced dynamic range (peak to noise floor). For these tests,

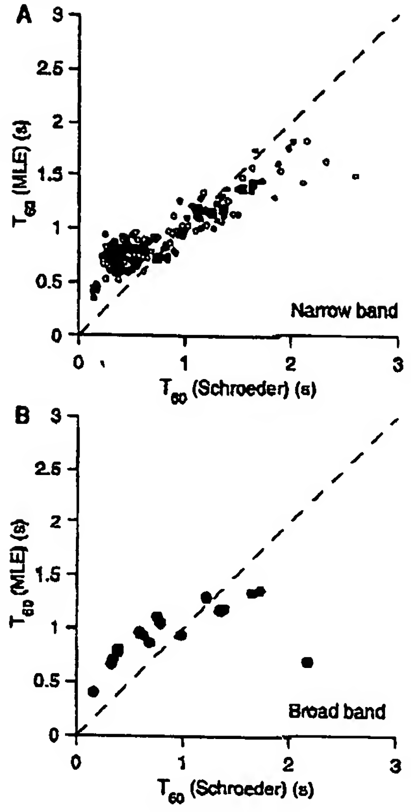

Figure 9: Reverberation time estimates from real environments. Seventeen tests in 12 environments were conducted using noise bursts. Decay time-constants were estimated using the MLE algorithm (ordinate) and the extrapolated T_{60} times were compared with estimates from Schroeder's method (abscissa). (A) Estimates of T_{60} in one-third octave bands with center frequencies exceeding 1 kHz (open circle) and their average (filled square). (B) Broad band estimates of T_{60} from the recorded room response. Averaged narrow band estimates are more accurate than broad band estimates, presumably due to the presence of low frequency components in the latter. Further, the MLE method over-estimates T_{60} (in comparison with Schroeder's method) when room reverberation is moderate ($<$ 0.3 s) whereas for higher values of T_{60} ($>$ 1.3 s) there is reasonable agreement. The single outlier is due to inaccuracies in the Schroeder estimate (see text for discussion). Results are from one presentation of noise burst in each test.

the Schroeder estimates of T_{60} (in seconds) were 2.18 (outlier), and 0.39, 0.39, and 0.33, respectively. The ML estimates, on the other hand, were 0.69 (outlier), 0.77, 0.8, and 0.67, respectively. Schroeder's method was inaccurate due to the reduced dynamic range. On the other hand, the ML estimates, while larger than the Schroeder estimates, were consistent and relatively robust to the reduction in dynamic range.

These results raise the issue of estimation in narrow bands. It appears, although it is by no means conclusive, that the upper one-third octave bands (over 1 kHz) may provide more accurate estimates than the lower bands. Since frequency decomposition is a standard part of most audio signal processing algorithms, such as found in beamforming, it may be useful to track estimates in the higher frequency bands, or in select bands where the energy is greatest. Tracking high-energy bands is likely to provide more temporal range in tracking decays before encountering the noise floor, and thus sharpen the peak in the histogram of estimates. Alternatively, averaging over all high-frequency bands can provide estimates that are in close agreement with T_{60} times obtained from Schroeder's method.

The findings suggest that there is good correlation between the estimates obtained from the MLE procedure and those obtained from Schroeder's method. While it is not possible to determine which method provides greater accuracy, we suggest that the values are correlated and in general agreement.

F. Estimation of RT from connected speech in real listening environments

The results presented in the preceding sections indicate that the ML estimator output is in good agreement with actual or simulated room RTs. In particular, the estimator can be applied to isolated word utterances, even though the naturally decaying offsets of terminating phonemes leads to an over-estimation of RT (see Fig. 6). Here, we test the performance of the procedure explicitly in a challenging estimation task, namely estimating room RT from connected speech.

A segment of speech (about 50 s in duration) from the Connected Speech Test (CST) corpus was played back in a partially open, circular foyer (one-third octave band analysis shown in Fig. 8, square symbols). The RT for this environment was first estimated with

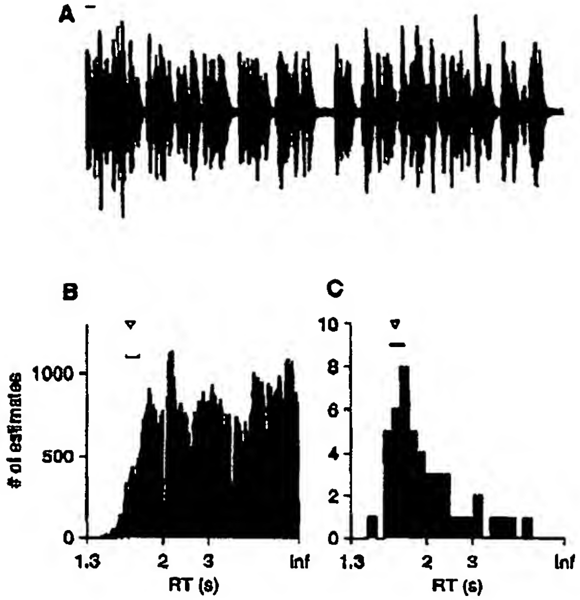

Figure 10: Evaluation of room reverberation time (RT) from connected speech played back in a partially open circular foyer. The RT for this environment as measured from hand-claps was 1.66 ± 0.07 s (Schroeder's method) and 1.62 s (from ML procedure). (A) Trace of CST passage (duration 50 s) recorded in the environment. Bar indicates 1 s. (B) The histogram of ML estimates over the duration of recording. The first peak in the aggregate histogram is the best RT estimate from connected speech (1.83 s). The horizontal bar is the range of RT estimates obtained from Schroeder's method, and the triangle indicates the ML estimate. (C) Peak values from histogram of estimates were obtained every 1 s, and the 50 peak values were used to produce the histogram shown. The best estimate of RT from this histogram is at the dominant peak (1.7 s), which is closer to the estimates obtained from pure decay curves. Thus, using short-term histograms as in (C) is more reliable than the long-term histogram shown in B. Overall, the results indicate that the ML estimator produces reliable estimates with connected speech.

hand-claps using Schroeder's method (1.66 ± 0.07 s) and independently confirmed with the ML procedure (estimated RT from histgoram was 1.62 s). The ML procedure was then applied to the recorded speech data (Fig. 10A). A histogram of room time constants for the duration of the recorded data was constructed (Fig. 10B). The order-statistics filter was used to select the first dominant peak in the histogram (RT = 1.83 s). This is the best RT estimate based on the aggregate data. It is possible to refine the procedure for arriving at the best estimate by applying the order-statistics filter at much shorter time intervals.

Towards this end, a histogram was constructed at intervals of 1 s, and the best RT estimate for this interval was obtained. The resulting best estimated from all one-second durations (50 in all) were binned to produce the histogram shown in Fig. 10C. It can be seen that the vast majority of the estimates were at RT = 1.7 s, which agrees with the mean value of 1.66 s from Schroeder's method (using hand-claps), and is well within its standard deviation (0.07 s; the one-sigma interval is indicated by the horizontal bar in Fig. 10B,C).

Given that terminal phonemes have a natural decay rate (see Fig. 6), it is not surprising that the ML procedure produces estimates somewhat larger than the real room RT. Further, the discrepancy between the actual RT and those estimated from connected speech arise from the absence of adequate numbers, and the limited duration, of gaps (see Fig. 10A). Thus, regions of free decay where estimation is accurate are limited. Notwithstanding these constraints, the procedure works well in part due to the decision-making capability built into the order-statistics filter. By selecting the first dominant peak (from the left) in the histogram, the filter in effect rejects spurious estimates and so reduces the error in the estimation procedure. The mean value of the histogram or its median, for instance, would result in significantly higher estimates of RT. The performance of the order-statistics filter can be further improved if one were to obtain a statistical characterization of gap duration from a large corpus of connected speech or other sounds. Such a characterization could provide a robust percentile cut-off value (see Eq. 17) which could then be used to select the best RT value for the room.

In conclusion, the ML procedure in combination with the order-statistics filter, provides a robust means for blind estimation of room RT. The procedure has been validated against Schroeder's method, and with real room data such as hand-claps, isolated word utterances, and connected speech.

V. DISCUSSION

A method for the blind estimation of reverberation time was presented. The method models the revereberant part of the room response, i.e., the diffusive tail, assuming that the response is an exponentially damped gaussian noise sequence. The input was assumed to be an ideal impulse. The objective was to determine the decay time-constant of the exponential. Using the statistical description of the process, a maximum-likelihood estimator (MLE) was derived, and solved numerically to arrive at the decay time-constant. This was then extrapolated to obtain the reverberation (T_{60}) time.

The estimation of reveberation time is a widely investigated problem. Traditionally, two approaches have been taken. The RT is computed analytically using formulae that incorporate the geometry and absorptive characteristics of the reflecting surfaces, or a test sound with known properties is radiated into the environment, and the RT is estimated from the received sounds. The former approach is embodied in the Sabine type formulae ([4, 9, 17, 20]; see [7, 23] for reviews), while the latter is based on Schroeder's decay curve analysis ([2, 18, 22]). In both approaches, prior knowledge of the environment or test sound is required. For example, in Sabine type formulae, the volume, surface area and absorption coeffiecients must be known; and in Schroeder's decay curve analysis, the test sound must be uncorrelated noise that is abruptly switched off at a known time, and followed by a sufficiently lengthy pause to track the decay. Thus, the methods are not suited for RT estimation from passively received sounds in unknown environments (i.e., blind estimation).

The MLE procedure removes these difficulties since it is based on a widely accepted model of the reverberant tail, namely the exponential decay model (see [23] for a discussion on how the Sabine type formulae are related to a linear decay of the sound pressure level after the source is turned off). Here, it is assumed that the amplitude of successive reflections are damped exponentially, while the fine structure is a random uncorrelated process. Since the model is a good approximation of reverberation in most diffusive environments, the method presented here provides a framework for blind estimation with wide applicability. The success of the approach also derives from the analytically tractable nature of the maximum-likelihood formulation, reducing the problem to the estimation of a single parameter that can be determined computationally. We also showed that for ongoing and onset segments of the sound, the estimates will assume implausible values since the model is not valid in these regions. However, an order-statistics filter downstream to the ML estimation can reject these estimates and extract the room RT with improved confidence. This is based on the intuitive idea that sounds cannot decay faster than the rate prescribed by the room time-constant, and thus, selecting the earliest peak improves the confidence of the estimates. To our knowledge, this approach has not been reported in the literature.

The two encouraging results of this study are the results obtained using speech sounds and the validation of the estimates using Schroeder's method. Speech sounds present particular problems to most estimation algorithms since they violate the two most commonly held assumptions, namely stationarity and Gaussian statistics. Further, even abruptly terminating phonemes such as stop consonants demonstrate a decay at the cessation of the sound. Such decays may be in the range of 5-40 ms and can increase the overall decay time estimated in reverberant environments. However, except for the increase in estimated decay time (a variation up to about 15% for sounds terminating in /d/) the tracking and histogram procedure works rather well, indicating that the method is relatively robust to model uncertainties.

Partially blind approaches to RT estimation are also available. 1) A neural network can be trained to learn the characteristics of room reverberation([3, 12]). Here, it is necessary to train the network whenever the environment changes. 2) The signal is explicitly segmented to identify gaps wherein decays can be tracked ([8]). It should be noted, that the order-statistics filter developed in this work performs an implicit segmentation of the signal by rejecting estimates that are implausible. 3) A blind dereverberation procedure can be used to obtain the room impulse response. However, the room impulse response must be minimum phase, a condition that most listening environments fail to satisfy ([10, 13]).

The ML procedure presented here is just one method for estimating room RT. Other methods are also possible. For instance the envelope of the sound can be extracted in the estimation interval, converted to sound pressure level, and a regression line could be fitted to obtain the T_{60} time. This is a blind version of the RT estimation procedure followed by Lebart et al. ([8]). The order-statistics filter can be applied to the histogram of estimates as with the ML procedure. The method is non-parametric and so is not subject to model uncertainties. This approach was used to estimate the decay rate of isolated word utterances (Fig. 6). Further work is needed to compare this procedure to the ML procedure.

The ML procedure is model-based and is expected to perform reasonably well in diffuse sound fields (i.e., uniform with respect to directional distribution) and where a single time-constant describes the reverberant tail. For most sound fields this is a resonable approximation (see [7] for a discussion on this point). The estimates of T_{60} are in reasonable agreement with Schroeder's method in most of the listening environments tested, including challenging situations where the source or recording microphone was close to a wall, or there was moderate background noise (see Fig. 7). Further, it provided consistent estimates even when the dynamic range of sound decay was reduced, in contrast to Schroeder's method which provides inaccurate estimates under these conditions (see Fig. 9 and accompanying text). While the ML procedure produces best results when there are isolated impulsive sounds or abruptly terminating white noise bursts, the results of tests with isolated word utterances and connected speech are in good agreement with the actual T_{60}. Thus, the procedure is expected to work under most listening conditions.

The method proposed here will perform poorly when there are room resonances and the sound pressure level decays nonlinearly with time. This can be a result of the room geometry, or positioning the recording microphone in a region of the sound field that is nondiffusive (e.g., against a reflecting surface). In addition to model failure, the performance of the estimator may be poor when there are insufficient numbers of gaps, or high levels of background noise. Good performance results when there are about 10% gaps and the peak sound level (at the time of offset) is about 25 dB SPL over the noise floor. Performance may also be poor when background noise is modulated (such as with background music or babble), since the procedure will attempt to track any modulation present in the environment and hence produce multi-modal histograms with peaks that may not be easily discriminated.

The blind estimation procedure suggested here can be applied in a number of situations. Since only passive sounds are used, any audio processor that has access to microphone input can estimate the room reverberation time, either in single-channel (broadband) or multi-channel (narrow-band) mode. One of the most interesting applications is in the selection of signal processing strategies tailored to specific listening environments. These include hearing aids and hands-free telephony. Most modern hearing aids have the ability to switch between several processing schemes depending on the listening environment. For instance, in highly diffusive environments, where the source to listener distance exceeds the critical distance, adaptive beamformers are ineffective. In this situation, it would be convenient to switch off the adaptive algorithm and revert to simple delay-and-sum beamforming. Such decisions can be made if there is a passive method for determining room reverberation characteristics. Other potential applications could include hands-free telephony, and sound level meters. A limitation of the method is its relatively poor performance with narrow-band signals whose center frequencies are below 1 kHz. However, the performance is good for broad-band signals, and narrow-band signals whose center frequncies exceed 1 kHz.

The computational costs of implementing the procedure are largely due to the iterative solution of the maximum-likelihood equation. We have developed fast algorithms for reducing the computational cost so that the procedure can be implemented in realtime (forthcoming publication). Thus, the method can be implemeted in passive listening devices to determine the reverberation characteristics of the environment.

[1] Bolt RK, MacDonald AD (1949). Theory of speech masking by reverberation. J. Acoust. Soc. Amer. 21: 577-580

[2] Chu WT (1978). Comparison of reverberation measurements using Schroeder's impulse method and decay curve averaging method. J. Acoust. Soc. Amer. 63(4): 1444-1450

[3] Cox TJ, Li F, Darlington P (2001). Extracting room reverberation time from speech using artificial neural networks. J. Audio Eng. Soc. 49(4): 219-230

[4] Eyring CF (1930). Reverberation time in "dead" rooms. J. Acoust. Soc. Amer. 1(2): 217-241

[5] Knudsen VO (1929). The hearing of speech in auditoriums. J. Acoust. Soc. Amer. 1: 56-82

[6] Kuttruff H (1995). A simple iteration scheme for the computation of decay constants in enclosures with diffusely reflecting boundaries.

[7] Kuttruff H (1991). Room Acoustics. London: Elsevier Science Publishers Ltd., Third Edition. J. Acoust. Soc. Amer. 98(1): 288-293

[8] Lebart K. Boucher JM, Denbigh PN (2001). A new method based on spectral subtraction for speech dereverberation. Acustics 87: 359-366

[9] Millington G (1932). A modified formula for reverberation. J. Acoust. Soc. Amer. 4(1): 69-82

[10] Miyoshi M, Kaneda Y (1988). Inverse filtering of room impulse response. IEEE Trans. Acoust. Speech and Sig. Proc. 36(2): 145-152

[11] Nabalek AK, Letowski TR, Tucker FM (1989). Reverberant overlap- and self-masking in consonant identification. J. Acoust. Soc. Amer. 86(4): 1259-1265

[12] Nannariello J, Fricke F (1999). The prediction of reverberation time using neural netowrk analysis. Appl. Acoust. 58: 305-325

[13] Neely ST, Allen JB (1979). Invertibility of room impulse response. J. Acoust. Soc. Amer. 66: 165-169

[14] Pitas I, Venetsanopoulos AN (1992). Order statistics in digital impage processing. Proc. IEEE 80(12): 1893-1921

[15] Poor V (1994). *An Introduction to Signal Detection and Estimation.* New York: Springer-Verlag

[16] Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992). *Numerical Recipes in C.* Cambridge: Cambridge Univ. Press

[17] Sabine WC. *Collected Papers on Acoustics.* Cambridge: Harvard University Press.

[18] Schroeder MR (1965). New method for measuring reverberation time. J. Acoust. Soc. Amer. 37: 409-412

[19] Schroeder MR (1966). Complementarity of sound buildup and decay. J. Acoust. Soc. Amer. 40(3): 549-551

[20] Sette WJ (1933). A new reverberation time formula. J. Acoust. Soc. Amer. 4(3): 193-210

[21] Tahara Y, Miyajima T (1998). A new approach to optimum reverberation time characteristics.

Appl. Acoust. 54: 113-129

[22] Xiang N (1995). Evaluation of reverberation times using a nonlinear regression approach. J. Acoust. Soc. Amer. 98(4): 2112-2121

[23] Young RW (1959). Sabine reverberation equation and sound power calculations. J. Acoust. Soc. Amer. 31(7): 912-921